# A Speaker Diarization System with Robust Speaker Localization and Voice Activity Detection

Yangyang Huang, Takuma Otsuka, and Hiroshi G. Okuno

**Abstract.** In real-world auditory scene analysis of human-robot interactions, three types of information are essential and need to be extracted from the observation data – who speaks *when* and *where*. We present a speaker diarization system that is used to accomplish the resolution. Multiple signal classification (MUSIC) is a powerful method for voice activity detection (VAD) and direction of arrival (DOA) estimation. We propose our system and compare its performance in VAD and DOA with the method based on MUSIC algorithm.

## 1 Introduction

Robot auditory functions are expected to facilitate an intuitive and natural human-robot interaction such as for the situation shown in Fig. 1. In such a situation, the ability from observations to recognize who speaks when and where is necessary. We deal with this problem as a speaker diarization task. A situation of free speech among multiple speakers is considered, which means speech without any scenario.

Speaker diarization is essential for various applications. In [1], a 3D auditory scene visualizer is proposed. It shows meta information, such as speech text and direction of speakers. They are extracted from observed signals by constructing a speaker diarization system using an audition system for robots [2]. In [3], a speaker diarization method including DOA and automatic speech recognition (ASR) is presented to estimate automatically "who speaks when and what" for a group meeting situation.

The speaker diarization described in this paper differs from most systems mentioned in [4] by the presence of overlap speech in input audio signals. To deal with the overlap speech, DOA estimation, VAD, sound source separation and source identification methods have been developed [2, 3, 5].

Yangyang Huang · Takuma Otsuka · Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University, Japan
e-mail: yangyang@kuis.kyoto-u.ac.jp

M. Ali et al. (Eds.): *Contemporary Challenges & Solutions in Applied AI*, SCI 489, pp. 77–82.
DOI: 10.1007/978-3-319-00651-2_11

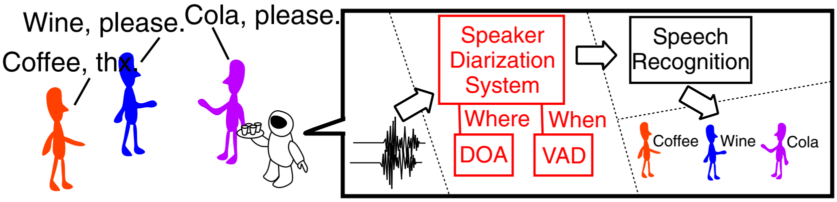


**Fig. 1** Need for speaker diarization. Robot waiter needs to understand who orders when and where to deliver the drinks

We focus on DOA and VAD, aiming to detect *when* and *where* a speech segment occurs. In particular, as the processing is done by connecting a plurality of elementary technology items in series, the performance of the preceding stage will affect the subsequent processing. Using robust methods in the preceding processing is required for various observations. For example, in the robot audition system, HARK, a multiple signal classification (MUSIC) algorithm is used for DOA and VAD in preceding processing, and source separation processing is followed. However, parameters for DOA and VAD need to be selected carefully, as they affect the performance of the entire system.

We also evaluate the DOA and VAD for free speech in a real environment. Ground truth for speaker direction and active speech segments are necessary, therefore We have corrected the ground truth data by using the MAC3D system. In addition, we use the rate of precision and recall for VAD evaluation.

Aiming to obtain higher accuracy for the speaker diarization system, We have improved the overall performance by using the independent vector analysis (IVA), which is a robust way of separating sound sources.

## 2 Problem Statement and System Configuration

For an input of multi-channel audio signals and output of speech segments and speaker direction, we assume that the transfer function of the microphone is known. A transfer function represents the transfer characteristic of the sound from each direction to the microphone array.

The proposed processing flow is shown in Fig. 2. After a short-time Fourier transform of the multi-channel audio signal, we apply IVA for separating the observed signals. Then VAD by threshold processing is used on the separated voice, and DOA by using the MUSIC algorithm.

### *2.1 Blind Source Separation by IVA*

The IVA method is an expansion of the blind source separation method, independent component analysis (ICA). This section will give an overview of ICA and then briefly describe the extension to IVA [6, 7].

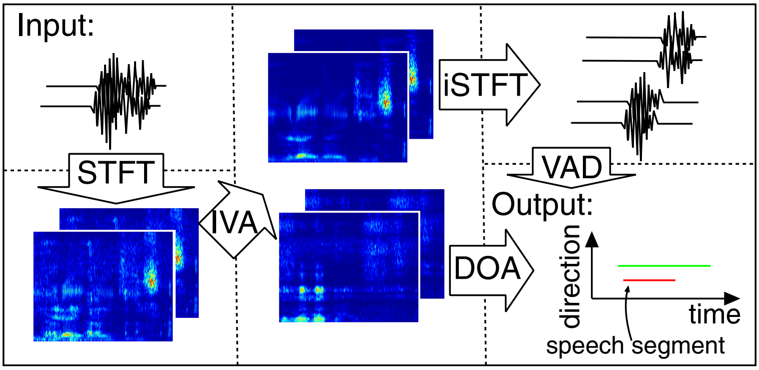


**Fig. 2** Processing flow of two channel audio signal for two-speaker free speech. Lines in time-direction coordinates represent speech segments by two speakers.

In the ICA method, the observed signal in the time-frequency domain $\mathbf{Z}_{t,f} = [z^1_{t,f}, ..., z^M_{t,f}]^T$ is modeled as: $\mathbf{Z}_{t,f} = \mathbf{A}_f \mathbf{Y}_{t,f}$.

Here, $\mathbf{Y}_{t,f} = [Y^1_{t,f}, ..., Y^M_{t,f}]^T$ is the audio signal for sound sources in the $t$th time frame and $f$th frequency bin, $\mathbf{A}_f$ is a mixing matrix. For the signal $\{\mathbf{Z}_{t,f}\}_{t=1}^T$, we calculate a separation matrix $\mathbf{W}_f$ which satisfies that components of $\hat{\mathbf{Y}}_{t,f}$ are statistically independent. $\hat{\mathbf{Y}}_{t,f} = \mathbf{W}_f \mathbf{Z}_{t,f}$.

However, the ICA method has a permutation problem: $\hat{\mathbf{Y}}_{t,f}$ calculated for each frequency bin $f$ may not in the same order as the original $\mathbf{Y}_{t,f}$. Thus, we need to select the correct component belonging to the same sound source in each frequency bin when restoring to the original audio signal.

By using this method, components of $\{\mathbf{Y}_{t,f}\}_{f=1}^F$ are deemed as an $F$-dimensional vector, this problem is avoided because $\{\mathbf{W}_f\}_{f=1}^F$ is optimized at the same time.

## *2.2 VAD by Threshold Processing*

We detect speech segment in separated audio signals. First, we restore the separated time-frequency domain signal $Y_{t,f}$ to the time domain waveform signal $y_t$ and extract one channel of them. Then we divide the waveform signal into shorter segment $\Delta t$ with no overlap. For each segment, the number of samples with absolute power larger than $T_v$ is calculated. A speech segment is confirmed if the number is more than $T_s$. The time-frequency domain signal corresponding to the determined speech segments are introduced into the next DOA process.

## *2.3 DOA by MUSIC Algorithm*

This section explains the MUSIC algorithm in general and discusses the advantages of using it directly for separated audio signals. The output of MUSIC is the spectrum which contains the energy calculated for each direction $\theta$ in each block $b$.

For time-frequency domain signals, the correlation matrix is calculated and averaged in a time block as: $\mathbf{R}_{b,f} = \sum_{t=(b-1)*\Delta T}^{b\Delta T} \mathbf{z}_{t,f}\mathbf{z}_{t,f}^{H}$, where H denotes a conjugate transpose operator. The eigenvalue decomposition of $R_{b,f}$ is given by $\mathbf{R}_{b,f} = \mathbf{E}_{b,f}\Lambda_{b,f}\mathbf{E}_{b,f}^{-1}$, where $\mathbf{E}_{b,f}$ is the eigenvector matrix and $\Lambda_{b,f}$ is the eigenvalue matrix. The eigenvector and eigenvalue can also be represented as $\lambda_{b,f,m}$ and $\mathbf{e}_{b,f,m}$.

The MUSIC spectrum is calculated as: $P_{b,f,\theta} = \frac{\|\mathbf{a}_{f,\theta}\mathbf{a}_{f,\theta}^{H}\|}{\sum_{m=N+1}^{M} |\mathbf{a}_{f,\theta}\mathbf{e}_{b,f,m}^{H}\|}$
where $a_{f,\theta}$ is the transfer function for the $\theta$th direction and $f$th frequency bin, and N denotes the number of sound source. A detailed explanation of MUSIC algorithm is provided in [8].

In the HARK system, Threshold processing is applied to the MUSIC spectrum calculated. The number of sound sources parameter and threshold parameter is used.

The use of IVA in our proposed system avoids the request for the parameter of the number source and threshold for the MUSIC spectrum.

# 3 Experiments

This section consists of three parts:

(1) Collection of reference data and create ground truth for free speech
(2) Evaluation criteria for VAD and DOA result
(3) Comparison of the baseline method and proposed method

## 3.1 *Reference Data and Ground Truth*

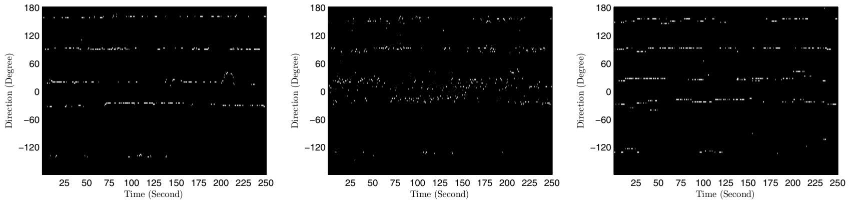
The ground truth for DOA and VAD takes a two-dimensional-array form represented as $x_{b,\theta}$. The indexes of the array denote the time block bins $b$ and the direction bins $\theta$. The value of $x$ denotes speaker ID or 0 for silence segment.

We obtained data of four free speech records and its reference data for evaluation. Each record was 240 seconds. The sampling rate was 16000 Hz, the block length for VAD was 0.5s, and the length of the STFT was 512 points.

## 3.2 *Evaluation Criteria*

DOA and VAD result is evaluated in frame wides. Impressively, how estimated result $\hat{x}_{b,\theta}$ is close to the ground truth $x_{b,\theta}$. Let the number of speech segments in ground truth by system be $S_a$, the number of speech segments in ground truth be $S_d$, the number of speech segments detected correctly be $S_c$, which is the difference between the correctly detected speech segment and the corresponding speech segment in ground truth. Evaluation criteria are defined as follows:

Precision: $R_p = \frac{S_c}{S_a}$, Recall: $R_r = \frac{S_c}{S_d}$, F measure: $F = \frac{2R_pR_r}{R_p+R_r}$.



(a) Ground truth (b) Baseline method (c) Proposed method

**Fig. 3** Comparison of DOA and VAD result

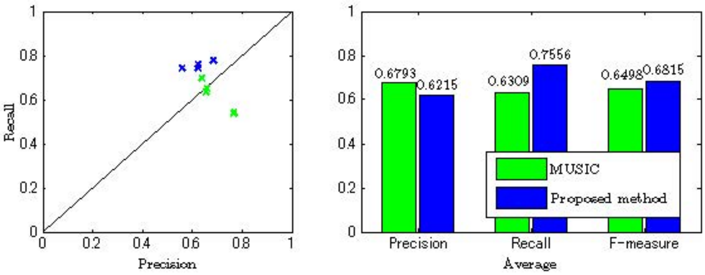


**Fig. 4** Quantitative result. Recall rate improved by over 10 percent. Green points in left figure show result by baseline method. Blue points shows result obtained by using the proposed method, with $T_v = 0.01$ and $T_s = 100$

### 3.3 *Results and Comparison*

In the IVA source separation process, the source number is 5, the number of speakers, and this number remains unchanged during the process. Threshold $T_v$ and $T_s$ for VAD is set to 0.01 (in 0∼1) and 100 (in 0∼8000).

Detection result for active speech segments are shown in Fig. 3. A comparison of the two methods is shown in Fig. 3. The proposed method has a higher F measure result and recall rate. Although precision rate is lower, clearly it can be improved by integrating other sensors.

## 4 Conclusion

We proposed a speaker diarization system which includes various processes connected in series. By using the IVA separation method as in the preceding process, the proposed system has the following advantages: (1) Higher recall and F measure than those of the baseline method. For a speaker diarization system, recall rate is the main requirement, which is equal to reduce the deletion error. (2) With the unchanged parameters, the proposed system still performs better than the baseline method in which parameters need to be fixed to improve performance.

**Acknowledgements.** This study was partially supported by a Grant-in-Aid for Scientific Research (S) 24220006.

## References

1. Kubota, Y., Yoshida, M., Komatani, K., Ogata, T., Okuno, H.G.: Design and implementation of 3d auditory scene visualizer towards auditory awareness with face tracking. In: Tenth IEEE International Symposium on Multimedia, pp. 468–476 (2008)
2. Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y., Tsujino, H.: Design and implementation of robot audition system 'hark' open source software for listening to three simultaneous speakers. Advanced Robotics 24(5-6), 739–761 (2010)
3. Araki, S., Hori, T., Fujimoto, M., Watanabe, S., Yoshioka, T., Nakatani, T., Nakamura, A.: Online meeting recognizer with multichannel speaker diarization. In: ASILOMAR, pp. 1697–1701 (2010)
4. Tranter, S.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing 14(5 ), 1557–1565 (2006)
5. Nakamura, K., Nakadai, K., Asano, F., Ince, G.: Intelligent sound source localization and its application to multimodal human tracking. In: Proceedings of the IEEE/RSJ International Conference on IROS, pp. 143–148 (2011)
6. Hyvarinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley Interscience (2001)
7. Ono, N.: Stable and fast update rules for independent vector analysis based on auxiliary function technique. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 189–192 (2011)
8. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation 34(3), 276–280 (1986)